CLAIMS

What is claimed is:

1. A system for document retrieval and/or indexing comprising:

a component that receives a captured image of at least a portion of a physical document; and

a search component that locates a match to the document, the search is performed over word-level topological properties of generated images, the generated images being images of at least a portion of one or more electronic documents.

2. The system of claim 1, further comprising a component that generates signature(s) corresponding to one or more of the generated images and generates a signature corresponding to the captured image of the document, the signatures identify the word-layout of the generated images, and the search performed *via* comparing the signatures of the generated images with the signature of the image of the captured document.

3. The system of claim 2, the signatures being at least one of hash tables and approximate hash tables.

4. The system of claim 3, the at least one of the hash tables and approximate hash tables comprising a key that is associated with a location and width of a word within at least one of the generated images and the image of the document.

5. The system of claim 2, further comprising a scoring component that assigns confidence scores corresponding to a subset of the generated images that are searched against.

6. The system of claim 5, wherein a generated image with the highest confidence score is selected as the match to the captured image of the document.

7. The system of claim 2, wherein the signature(s) corresponding to the one or more generated images comprise a tolerance for error.

8. The system of claim 2, wherein a portion of the signature(s) associated with the one or more generated images is compared to a corresponding portion of the signature of the image of the captured document.

9. The system of claim 8, wherein the signature(s) corresponding to the one or more generated images that have a threshold number of matches to the corresponding portion of the signature of the captured image of the document are retained for further consideration.

10. The system of claim 9, further comprising a component that assigns confidence scores when a threshold number of signatures are being retained for further consideration.

11. The system of claim 2, the signatures corresponding to the one or more generated images and the signature of the image of the captured document are generated at least in part upon a location of at least a portion of each word in the generated images and the image of the captured document, respectively.

12. The system of claim 11, the signatures corresponding to the one or more generated images and the signature of the captured image of the document further generated at least in part upon a width of each word in the captured image and the generated images, respectively.

13. The system of claim 2, further comprising:

a component that generates tree representations related to the generated images and the captured image of the document, the tree representations being a hierarchical representation of the generated images and the captured image of the document, wherein the tree representations convey which segments of the generated images and which segments of the image of the documents include a word; and

a comparison component that compares a tree representation related to the generated images with the tree representation related to the captured image of the document.

14. The system of claim 1, further comprising a component that reduces noise in the captured image of the document.

15. The system of claim 1, further comprising a component that generates a grayscale image of the captured image of the document.

16. The system of claim 1, further comprising a connecting component that connects characters within a word of the generated images and the captured image without connecting words of the generated images and the captured image.

17. The system of claim 16, the generated images and the captured image being binary images, the connecting component performs a pixel dilation of the binary images.

18. The system of claim 17, the connecting component alters resolution of the captured image of the document to facilitate connecting characters within a word of the captured image of the document without connecting disparate words within the captured image of the document.

19. The system of claim 1, further comprising a caching component that automatically generates an image of an electronic document at a time such electronic document is printed.

20. The system of claim 19, further comprising an artificial intelligence component that infers which printed documents should have associated stored images.

21. The system of claim 1, further comprising an artificial intelligence component that excludes a subset of the generated images from the search based at least in part upon one of user state, user context, and user history.

22. The system of claim 1, wherein at least one of the generated images is associated with an entry within a data store, the entry comprising one or more of an image of a page of an electronic document and a signature that identifies the image of the page, the signature based at least in part upon topological properties of words within the image of the page.

23. The system of claim 22, the one or more of the image of the page of the electronic document and the signature that identifies the image of the page associated with one or more of a URL that identifies a location of the electronic document, the electronic document, a hierarchical tree representation of the image of the page of the electronic document, OCR of the image of the page, data relating to a number of times the image of the page has been accessed, customer records, payment information, and workflow information.

24. A method that facilitates indexing and/or retrieval of a document, comprising:

generating a plurality of images of electronic documents, at least one of the images of electronic documents corresponding to a printed document;

capturing an image of a printed document after such document has been printed;

receiving a query requesting retrieval of an electronic document corresponding to the image of the printed document;

generating one or more signatures corresponding to at least a portion of one or more of the generated images, the signatures generated at least in part upon word-layout within the image(s);

generating a signature corresponding to at least a portion of the captured image, the signature generated at least in part upon word-layout within the captured image; and

comparing the one or more signatures corresponding to the one or more generated images to the signature corresponding to the captured image.

25. A method that facilitates indexing and/or retrieval of a document, comprising:

receiving a captured image of at least a portion of a document; and

searching data store(s) for an electronic document corresponding to the captured image, the search performed *via* comparing topological word properties within the captured image with topological word properties of generated images corresponding to a plurality of electronic documents.

26. The method of claim 25, further comprising:

generating signatures corresponding to the generated images, the signatures based at least in part upon location and width of each word within the generated images;

generating a signature corresponding to the captured image of the document, the signature based at least in part upon location and width of each word within the captured image; and

comparing the signatures corresponding to the generated images with the signature corresponding to the captured image of the document.

27. The method of claim 25, further comprising:

partitioning the captured image of the document into a plurality of segments;

partitioning the generated images into segments substantially similar to the segments of the captured image of the document; and

comparing the word layout of the captured image of the document with the word layout of the generated images only within corresponding segments of the captured image of the document and the images within the data store(s).

28. The method of claim 27, further comprising:

assigning confidence scores to the signatures corresponding to the generated images based at least in part upon a similarity between the word layout of the captured image and the word layout of the generated images.

29. The method of claim 25, further comprising:

partitioning the captured image of the document to create a hierarchy of segments;

partitioning the generated images to create a hierarchy of segments corresponding to the hierarchy of segments related to the captured image of the document;

assigning the segments in the captured image of the documents and the segments in the generated images a first value when the segments comprise a word;

assigning the segments in the captured image of the documents and the segments in the generated images a second value when the segments do not comprise a word;

comparing the hierarchy of segments; and

removing one or more generated images from consideration when a segment associated with the one or more generated images assigned the second value and a corresponding segment associated with the captured image of the document is assigned the first value.

30. The method of claim 25, further comprising reducing noise in the captured image of the document prior to searching the data store(s).

31. The method of claim 30, wherein reducing noise comprises one or more of:

providing a filter that removes markings that have a width greater than a threshold width;

providing a filter that removes markings with a width less than a threshold width;

providing a filter that removes markings with a height greater than a threshold height; and

providing a filter that removes marking with a height less than a threshold height.

32. The method of claim 25, further comprising generating a grayscale image of the captured image of the document prior to searching the data store(s).

33. A system for indexing and/or retrieval of a document, comprising:

means for generating an image of an electronic document when the electronic document is printed;

means for capturing an image of the document after the document has been printed;

means for retrieving the electronic document, the means based at least in part upon comparing location and width of words within the captured image to the location and width of words within the generated image.

34. The system of claim 33, further comprising:

means for generating a signature that includes features that are highly specific to the generated image; and

means for generating a signature corresponding to the captured image, the signature includes features that are highly specific to the captured image.

35. The system of claim 34, further comprising means for comparing the signature corresponding to the generated image with the signature corresponding to the captured image.

36. The system of claim 34, further comprising means for accounting for error that occurs when capturing the image of the printed document.

37. The system of claim 33, further comprising:

means for partitioning the generated image into a plurality of segments;

means for partitioning the captured image into a plurality of substantially similar segments; and

means for comparing a segment of the stored image with a corresponding segment of the captured image.

38. A system that facilitates indexing and/or retrieval of a document, comprising:

a query component that receives an image of a printed document;

a caching component that generates and stores an image corresponding to the image of the document prior to the query component receiving the image of the printed document; and

a comparison component that retrieves the stored image *via* comparing at least one of location and width of words within the stored image to location and width of words within the image of the printed document.

39. A computer readable medium having computer executable instructions stored thereon to return stored image(s) of an electronic document to a user based at least in part upon topological word properties of captured image(s) corresponding to the printed document.

40. A computer readable medium having a data structure thereon, the data structure comprising:

a component that receives image(s) of at least a portion of a printed document; and

a search component that facilitates retrieval of an electronic document, the electronic document corresponding to the image(s) of the printed document, the retrieval based at least in part upon similar word-level topological properties when comparing the image(s) of the printed document and generated image(s) of the electronic document.

41. A personal digital assistant comprising the system of claim 1.

42. A signal having one or more data packets that facilitate indexing and/or retrieval of a document, comprising:

a request for retrieval of a stored image of at least a portion of an electronic document;

a signature of an electronic image of a printed document corresponding to a signature of the images of the requested stored electronic document, the signatures based at least in part upon word layout of the images; and

a component that facilitates comparison of the signature of the image of the printed document with the signature of the image of the requested stored document.